Discourse Salience and Pronoun Resolution in Hindi

Rashrni Prasad
University of Pennsylvania, rjprasad@babel.ling.upenn.edu

Michael Strube
European Media Laboratory GmbH, Michael.Strube@eml.villa-bosch.de

Discourse Salience and Pronoun Resolution in Hindi

Discourse Salience and Pronoun Resolution in Hindi*

Rashmi Prasad and Michael Strube

1 Introduction

This paper investigates anaphoric reference in Hindi, with particular focus on the use and interpretation of third person personal pronouns to realize anaphoric relationships between noun phrases. We have two specific goals. The first is inspired by the central idea of Centering theory (Grosz et al. 1995), namely, that each utterance in a discourse evokes certain discourse entities (Webber 1978; Prince 1981) which comprise the list of *forward-looking centers* (the Cf-list), in Centering terms, and which are ranked according to their salience. The anaphoric relationships in the local discourse segment (Grosz and Sidner 1986) are dependent on the Cf-list ranking, in that the more highly ranked entities in an utterance are more likely to be talked about in the following utterance. Investigation of the factors that determine the Cf-list ranking—which have not yet been completely specified—has, therefore, constituted an important aspect of the research for Centering theory in particular, and for discourse anaphora in general. Furthermore, crosslinguistic research has revealed that this ranking is dependent on language specific factors (Walker et al. 1994; Turan 1995; Strube and Hahn 1999, among others). Our purpose here is to investigate such factors in Hindi, with special focus on the role of grammatical function, word order, and information status. We also propose a novel, general method for determining these ranking factors.

Centering theory has also guided the development of pronoun resolution algorithms, such as the BFP algorithm and the algorithm developed by Strube (1998, henceforth, S-list algorithm). Both algorithms regard the notion of relative salience to be crucial for the resolution of pronouns, and in order to apply these algorithms for pronoun resolution in any language, the first task is to be able to determine the Cf-list ranking criteria for that language. Our second goal, therefore, is to apply these algorithms to the resolution of pronouns in Hindi texts by incorporating the results of our analysis of relative salience in Hindi. In doing so, we show that the BFP algorithm cannot be successfully implemented for pronoun resolution in Hindi and that, in fact, the same prob-

*We would like to thank Jennifer Arnold, Miriam Eckert, Aravind Joshi, Kathy McCoy, Ellen Prince, and an anonymous reviewer for their invaluable comments. This work was partially funded by a post-doctoral fellowship from IRCS (NSF SBR 8920230).

lems extend straightforwardly to an implementation of this algorithm in any other language. We argue that better results can be obtained with an algorithm that does not use the Centering notions of the *backward-looking center* and the *centering transitions* for the computation of pronominal antecedents, such as the S-list algorithm proposed by Strube (1998).

Section 2 presents a brief overview of Centering theory. In Section 3, we present our method for determining relative salience, and show that *grammatical function* is a crucial factor for ranking discourse entities in Hindi, with word order and information status having no independent effect on salience. In Section 4, we present the BFP algorithm and discuss the problems that it presents for pronoun resolution, using examples from Hindi as an illustration. In Section 5, we describe the S-list algorithm and adapt it to results obtained for Hindi. Finally, in Section 6, we compare the performance of the two algorithms for the resolution of pronouns in Hindi texts.

2 Centering Theory

Centering theory is a model of local discourse coherence which makes predictions about the inference load placed on a hearer in processing a discourse segment. The crucial claims of the theory are as follows:

- Discourses are composed of constituent segments, each one of which consists of particular utterances.
- Each utterance U_i in a given discourse segment is assigned a list of *forward-looking centers*, $Cf(U_i)$, where *centers* are semantic entities in the discourse model (Webber 1978).
- Each utterance (other than the segment-initial utterance) is assigned a unique *backward-looking center*, $Cb(U_i)$.[1]
- The list of *forward-looking centers*, $Cf(U_i)$, is ranked according to discourse salience, with the highest ranked element of $Cf(U_i)$ being called the *preferred center*, $Cp(U_i)$ (Brennan et al. 1987).
- The most highly ranked element of $Cf(U_{i-1})$ that is *realized* in U_i is the $Cb(U_i)$.[2]

[1]The *Cb* corresponds to the discourse entity that the utterance is most centrally about, and is similar to the notion of the *topic* (Reinhart 1981; Horn 1986).

[2]An utterance U *realizes* a center c if c is an element of the situation described

The theory defines transition relations across pairs of adjacent utterances (see Table 1, taken from Walker et al. (1994)). The transitions differ from each other according to (a) whether Cb's of successive utterances are equal or not, and (b) whether the Cb of any utterance corresponds to the Cp of that utterance or not.

	$Cb(U_i) = Cb(U_{i-1})$ OR $Cb(U_{i-1}) = [?]$	$Cb(U_i) \neq Cb(U_{i-1})$
$Cb(U_i) = Cp(U_i)$	CONTINUE	SMOOTH-SHIFT
$Cb(U_i) \neq Cp(U_i)$	RETAIN	ROUGH-SHIFT

Table 1: Transition Types

The theory also proposes two rules, violations of which are predicted to increase the hearer's inference load for the interpreting the discourse segment.

Rules: For each utterance, U_i, in a discourse segment $U_1, \ldots, U_m$:

1 If some element of $Cf(U_{i-1})$ is realized as a pronoun in U_i, then so is the $Cb(U_i)$.

2 Transition sequences are ordered. CONTINUE > RETAIN > SMOOTH-SHIFT > ROUGH-SHIFT.

One indeterminate part of the theory is the manner in which the Cf-list is ranked. The ranking plays a crucial role as it determines which of the elements of $Cf(U_{i-1})$ realized in U_i will be the $Cb(U_i)$, upon which depends the calculation of the transitions across adjacent utterances and thus of the inference load for interpretation.

3 Relative Salience in Hindi

Crosslinguistic research within the framework of Centering theory has led to the speculation that languages may vary with respect to which linguistic properties affect the salience of discourse entities.[3] For instance, Brennan

by U, or c is the interpretation of some subpart of U; a center is *directly realized* if it corresponds to a phrase in an utterance (Grosz et al. 1995).

[3]For details on cross-linguistic research on Centering, see Sidner (1979), Gordon et al. (1993), Grosz et al. (1995) for English; Di Eugenio (1998) for Italian; Prince

et al. (1987) assume the following ranking for the Cf-list in English: *subject > object > object2 > other subcategorized functions > adjuncts.* Walker et al. (1994) extend the *Cf*-ranking criteria for Japanese in order to account for zero-pronouns, topic-marked NPs and NPs which are emphasized by empathy-marked verbs. They propose the following ranking for Japanese: *topic (grammatical/zero) > empathy > subject > object > other(s).* Rambow (1993) and Strube and Hahn (1999) suggest that the ranking in German might follow the *surface order position.* Gordon et al. (1993) suggest that *sentence-initial position* seems to contribute to salience. Turan (1995) argues that the Cf-list ranking in Turkish is associated with either *grammatical relation* or a *semantic role* hierarchy, and also provides evidence to show that word order does not play a role. Strube and Hahn (1999) propose that the ranking criteria for the Cf-list in German is partly determined by the information status of the discourse entities. They distinguish between *old*, *mediated*, and *new* discourse entities, and propose the following ranking: *old > mediated > new.*[4]

In the following section, we present a novel, general method for determining which aspects of linguistic knowledge play a role in ranking the elements of the Cf-list. We apply this method to Hindi and discuss the influence of grammatical function, word order, and information status.[5]

3.1 Method for Determining Relative Salience

Our method for determining relative salience invokes Rule 1 of Centering theory.[6] According to this rule, if anything is pronominalized in an utterance, the Cb must be, too. In other words, if there is a single pronoun in an utterance U_i, it must be the Cb of U_i and it must cospecify with the highest ranked entity among those in U_{i-1} *that are realized* in U_i.[7]

(1994) for Yiddish; Kameyama (1985), Walker et al. (1994) for Japanese; Hoffman (1998), Turan (1995) for Turkish; Rambow (1993), Strube and Hahn (1999) for German; and Dimitriadis (1996) for Greek.

[4]The information status distinctions in Strube and Hahn (1999) correspond to Prince's (1981) distinctions in the following manner: *old* entities correspond to *unused* and *evoked* entities, *new* entities correspond to *brand-new* entities, and *mediated* entities correspond to *inferrables*, *containing inferrables* and *anchored brand-new* entities.

[5]In this paper, we ignore other factors that have been argued to affect Cf-list ranking, such as lexical semantics, intonation, tense etc..

[6]Rule 1 captures the intuition, originally stated in Sidner (1979, 1981), that pronominalization is one of the markers of salience (*immediate focus* in Sidner's terms).

[7]We assume that Rule 1 (as well as the other rules and constraints of Centering theory) has some cognitive reality (Gordon et al. 1993; Hudson-D'Zmura and Tanenhaus

We searched our corpus for utterance *pairs*, U_{i-1} and U_i, which satisfy the following three conditions:

1. U_i, realizes *only two* of the entities from U_{i-1}.
2. In U_i, only one of the NPs realizing these entities is pronominalized.
3. The pronoun in U_i is ambiguous (for gender and number) between the two entities in U_{i-1}.[8]

The procedure for determining the relative salience of entities in any utterance U_{i-1} is as follows: given Rule 1 and the conditions stated above, if two discourse entities X and Y in U_{i-1} are both realized in U_i, with only Y being realized as a pronoun (in U_i), then Y must be the Cb of U_i and must cospecify with the highest ranked of all the entities in U_{i-1} that are realized in U_i. Since X and Y are the only two entities in U_{i-1} realized in U_i, Y must be ranked higher than X (or be more salient than X). Conversely, if it is X (and not Y) that is realized as a pronoun in U_i, then by the same reasoning X must be more salient than Y in U_{i-1}.

The method described above was applied to a corpus consisting of short stories. The 560 utterance pairs that filled the defined criteria were further categorized in different groups according to the linguistic properties of the NPs, such as grammatical function, word order and information status. Within each of these groups, further subgroupings were done according to the pair of factors that were being compared for relative salience. For example, one such grouping was in terms of grammatical function, and this had further subgroups—one for comparing the salience of subjects and direct objects, another for comparing the salience of direct objects and indirect objects, and so on.

Note, however, that Rule 1 is *not* non-violable. In fact, the calculation of discourse coherence in Centering theory is partly based on the assumption that speakers can be expected to violate Rule 1. However, for the task of determining relative salience according to our method, such an expectation seems to create the following *determinacy* problem. Consider any group of n utterance pairs, U_{i-1} and U_i, such that two entities X and Y in U_{i-1} in all pairs exhibit the linguistic properties characterizing the group, with X having the property LX and Y having the property LY. As was described above,

1998).

[8] We selected utterance pairs with ambiguous pronouns to reduce the noise resulting from Rule 1 violations that are induced by the availability of grammatical information that allows inferential (defeasible) reasoning on the part of the hearer.

both these entities are realized in U_i in all pairs with only one of them being realized as a pronoun. Now, if all the n pairs in the group pronominalize the entity with the same property, i.e., either LX or LY, then the relative salience of the entities X and Y in U_{i-1} is completely determinate. However, if k pairs pronominalize the entity with property LX and $n - k$ pairs pronominalize the entity with property LY, then the analyst faces the question of deciding which one of the sets of pairs is the true indicator of salience. These two opposing cases are illustrated in the schematic diagram in Figure. 1, where U_{i-1} and U_i are adjacent utterances, W, X, Y, and Z are the discourse entities in U_{i-1}, and A, B, X, and Y are the discourse entities in U_i. Both cases realize only two entities from U_{i-1} in U_i, namely X and Y, and X and Y in U_{i-1} have either the linguistic property LX or LY. In U_i in Case 1, Y is realized as a pronoun (labeled 'pro') and X as a full noun phrase (labeled 'NP'), whereas Case 2 shows the opposite situation. The task, therefore, is to decide which of

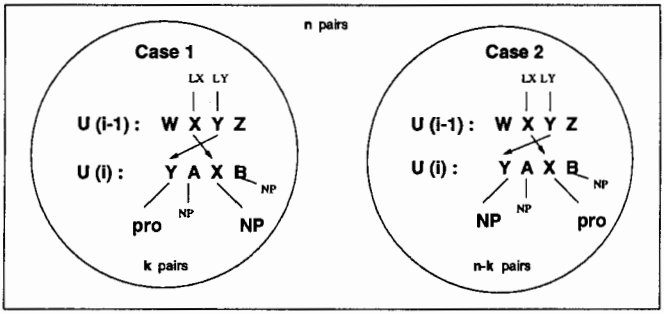

Figure 1: Opposing behaviors for Salience

the cases is a true indicator of relative salience and which constitutes a Rule 1 violation. Our decision is based on frequency of occurrence of the two cases in the corpus, in that the one which occurs with greater frequency is taken to be the indicator of salience. This is motivated by the assumption that speakers exhibit a preference for adhering to Rule 1 and that this preference can be observed in naturally occurring discourse in terms of greater (given that the opposing case does occur at all) frequencies of occurrence (Jaspars and Kameyama 1995). As will be seen later, this assumption is empirically justified in our corpus.

3.2 Some Facts about Hindi

Before proceeding with the investigation of the factors determining relative salience in Hindi, a few remarks about the language are in order. The subject-indirect object-direct object-verb (S-IO-DO-V) order is the default word order in Hindi. However, the language allows many other word orders (example (1)) which signal distinctions in meaning (Gambhir 1981) relating to information structure (Vallduví 1990). Hindi has a rich case system, though case marking is not obligatory. Pronouns are unmarked for gender and only partially marked for number.[9] In particular, though some forms, like *usne* 'he', *usko* 'him', are unambiguously singular, some forms can be both singular and plural, like *unhone* 'he/they', or *unko* 'him/them'.

(1) a. malay-ne sameer-ko kitaab dii (S-IO-DO-V)
malay-ERG sameer-DAT book-ACC gave
'Malay gave the book to Sameer'

b. malay-ne kitaab sameer-ko dii (S-DO-IO-V)

c. sameer-ko malay-ne kitaab dii (IO-S-DO-V)

d. sameer-ko kitaab malay-ne dii (IO-DO-S-V)

e. kitaab malay-ne sameer-ko dii (DO-S-IO-V)

f. kitaab sameer-ko malay-ne dii (DO-IO-S-V)

Hindi has verb agreement with the subject or the direct object. The agreement inflection is marked for person, number, and gender. With respect to the form and information status of noun phrases, Hindi has (non-obligatory) definite and (obligatory) indefinite articles. Following Prince (1992), the NPs with the indefinite article usually refer to hearer-new and discourse-new entities, whereas NPs with the null/overt definite article usually refer to hearer-old and/or discourse-old entities.

3.3 Factors Determining the Ranking

In all the examples in this section, the pronouns and the NPs they cospecify with are indicated in boldface and by coindexation.[10] In each case, the pronoun is ambiguous with respect to the person, number or gender features of

[9]Hindi also has zero pronouns, but their occurrence is heavily constrained, unlike in Italian (Jaeggli and Safir 1989) or Japanese (Kameyama 1985). In this paper, we do not investigate the interpretation of null pronouns in Hindi.

[10]To avoid confusion between the ambiguous denotation of the Hindi pronouns and the unambiguous English translations, the pronouns are glossed as **pro**.

the two entities whose salience is being compared (and which are indicated by square brackets with grammatical category labels).[11]

3.3.1 Grammatical Function: Subject vs. Direct Object/Indirect Object

Example (2) illustrates that the subject is ranked higher than the direct object. Both the subject and direct object in (2a) are realized in (2b), but it is the subject that is realized as the pronoun. The subject, therefore, qualifies as the *Cb* of (2b) and consequently as the more highly ranked of the two entities in (2a) that are realized in (2b).

(2) a. aise maukoN par **[$_S$ savaariyaan]$_i$** [$_{DO}$ chaate]$_j$ taan letii
such occasions on **[$_S$ passengers]$_i$** [$_{DO}$ umbrellas]$_j$ open take
haiN
3pl.fem.prs
'On such occasions the passengers open umbrellas'

b. kabhi-kabhi tej havaa se [chaate]$_j$ **[unke]$_i$** haath
sometimes fast wind with [umbrellas]$_j$ **[pro-POSS]$_i$** hands
se urr bhii jaate haiN
from fly also go 3pl.fem.prs
'Sometimes, because of the strong winds, the umbrellas even fly away from their hands'

By the same argument, example (3) shows that the subject is ranked higher than the object within the prepositional argument of the verb. Both the subject as well as the prepositional object in (3a) are realized in (3b), but it is the subject that is pronominalized and therefore, it qualifies as the *Cb* of (3b) and as more highly ranked than the prepositional object in (3a).

(3) a. kuch der pashchaat, **[$_S$ ek shramik]$_i$** [$_{PP}$ [$_{PO}$ us yuvak]$_j$
some time after, **[$_S$ a laborer]$_i$** [$_{PP}$ [$_{PO}$ that youth]$_j$
ke paas] aayaa
near to] came
'After some time, a laborer came up to the youth'

b. **[usne]$_i$** [yuvak]$_j$ se puuchaa ki "kyaa aagyaa hai?"
[pro-ERG]$_i$ [youth]$_j$ of asked that "what wish is?"
'He asked the youth, "what is your wish?"'

[11] S = subject, DO = direct object, IO = indirect object, PP = prepositional phrase, PO = prepositional object, ACC = accusative, ERG = ergative, DAT = dative.

334 utterance pairs in the corpus consisted of the subject and either the direct object, or the indirect object, or the prepositional complement being realized in both the utterances in the pair (see Table 2). 322 cases show the subject ranked higher (in particular, out of 149 pairs for comparing the subject and the direct object, 144 have the subject realized as a pronoun in the second utterance (96%) and 5 have the direct object as the pronoun (4%); out of 57 pairs for comparing the subject and the indirect object, 50 have the subject realized as a pronoun (87%) and 7 have the indirect object realized as the pronoun (13%); finally, out of 128 pairs for comparing the subject and the prepositional complement, all have the subject realized as the pronoun (100%)).

3.3.2 Grammatical Function: Direct Object vs. Indirect Object

Examples like (4) suggest that a higher degree of salience is attributed to entities denoted by the direct object when compared to indirect objects. Both the direct object and indirect object in (4a) are realized in (4b), but it is the direct object that is pronominalized. Note that the pronoun in (4b) is unmarked for gender and number and is therefore ambiguous between the two antecedents in (4a) (DO is masculine and IO is feminine).

(4) a. $[_S$ dukaandaar ne$]_i$ $[_{DO}$ **kaii namune ke kapDe**$]_j$ $[_{IO}$ un
$[_S$ shopkeeper ERG$]_i$ $[_{DO}$ **many types of clothes**$]_j$ $[_{IO}$ those
striiyoN ko$]_k$ dikhaaye
women ACC$]_k$ show-3sg.m.pst
'The shopkeeper showed many types of clothes to those women'

b. [un striiyoN ko$]_k$ **[unme]**$_j$ se kuch pasand aaye
[those women ACC$]_k$ **[pro]**$_j$ of some like come-3pl.pst
aur kuch unhone alag hataa diyaa
and some they-ERG aside remove give.3sg.pst
'The women liked some of them and some they removed aside'

The corpus contains 22 pairs illustrating the comparison above (see Table 2), and all of them have the direct object realized as the pronoun in the second utterance.

Other ranking comparisons constitute the rest of the pairs in the corpus, i.e., 204 pairs (see Table 2 for detailed figures). The partial ranking with respect to the grammatical functions that we were able to investigate can thus be specified as follows: *subject* > *direct object* > *indirect object/PP object* > *adjuncts*. In addition, we also specify that for a possessive noun phrase, *possessor* > *head noun*.

Ranking	Number	Total	Frequency
Subject > Direct Object	144	149	96
Subject > Indirect Object	50	57	87
Subject > PP Object	128	128	100
Direct Object > Indirect/PP Object	22	22	100
Subject/Object> Adjunct	96	110	87
Possessor > Head	22	22	100
Subject > Possessor of Direct Object	50	50	100
Indirect Object > Possessor of Subject	22	22	100
Total	534	560	95

Table 2: Frequencies for Relative Salience of Grammatical Functions

3.3.3 Against Word Order and Information Status

The surface order of constituents has been argued to be a determining factor for relative salience in German (Rambow 1993; Strube and Hahn 1999). Our Hindi corpus does contain utterance pairs in which the entity represented by the sentence-initial constituent is found to be the *Cb* in the next utterance, but in fact, these constituents are always the subject. Furthermore, there are also cases in which the subject occurs in some non sentence-initial position, and examples such as (5a-b) show that it is the subject, rather than the sentence-initial constituent occurring in the initial position, which is realized as the *Cb* in the following utterance.

(5) a. $[_{PO}$ bailon$]_i$ ke biich $[_S$ **ek purush**$]_j$ kharaa hai
$[_{PO}$ buffalos$]_i$ of between $[_S$ **a man**$]_j$ stand.3sg 3sg.pres
'There is a man standing between the buffalos'

b. **[vah]**$_j$ $[_{IO}$ in bailon ko$]_i$ charaa daal rahaa hai
[he]$_j$ $[_{IO}$ these buffalos DAT$]_i$ fodder put do.3sg 3sg.pres
'He is giving fodder to these buffalos'

In the transition from (5a) to (5b), the subject is not in sentence-initial position in (5a), but still denotes the centered entity, since it is the antecedent of the pronoun, the *Cb*, in (5b). The example also shows that unlike German, the information status of discourse entities does not play a role in the Cf-list ranking in Hindi. In (5a), a new entity, *ek purush*, 'a man' is introduced, and the utterance also contains a discourse old entity, *bailon*, 'buffalos'. Ranking the entities according to the critieria suggested for German (with *old* > *mediated* > *new* (Strube and Hahn 1999)) cannot account for the realization of

the discourse new entity, *a man*, as a pronoun, the Cb, in (5b). What really matters is that the discourse new entity is found to be in the subject position of the sentence. In all such cases, where some NP denoting a discourse-old entity is preposed to the sentence-initial position, it is the discourse new entity in subject position that is pronominalized in the following utterance.

In conclusion, the corpus revealed an overwhelming influence of grammatical function on the salience of the discourse entities evoked in an utterance, and therefore of the Cf-list in Centering terms. Furthermore, word order and information status do not seem to play any independent role in determining relative salience in Hindi.

4 The BFP Algorithm

In this section, we apply the BFP algorithm for pronoun resolution in Hindi using the ranking obtained in the previous section. The algorithm described by Brennan et al. (1987) incorporates the centering rules and transitions and consists of three basic steps (as described by Walker et al. (1994)).[12]

1. GENERATE possible *Cb-Cf* combinations (anchors).
2. FILTER by constraints, e.g., contra-indexing, sortal predicates, centering rules and constraints.[13]
3. RANK by transition orderings.

In applying the BFP algorithm, we found that the algorithm makes two types of strategic errors. The first is caused by its preference for *Continue* transitions. This preference implies that a pronoun in U_i is more likely to refer to the $Cb(U_{i-1})$ than to the $Cp(U_{i-1})$ when $Cb(U_{i-1}) \neq Cp(U_{i-1})$ (= *Retain* or *Rough-shift*). This preference does not hold for Hindi, as shown in example (6).[14] The tables below each utterance contain the filtered anchors for that utterance. The second column in the tables represents the discourse entities and the third column represents the corresponding surface expressions.

(6) a. Congress adhyaksh$_1$ *unse*$_2$ aise mile
Congress director$_1$ him$_2$-with this-way met
'The Congress director met him in such a way,'

[12]The algorithm has also been applied to Japanese by Walker et al. (1994).

[13]Contra-indexing constraints on coreference originate from Binding theory (Chomsky 1981).

[14]From "Bihari Babu ke haseen sapne". Article in *India Today*. Issue: 31 December 1997.

(i)	Cb :	Laalu : (2)	*unse*
	Cf :	[Director (1) :	*Congres adhyaksh*
		Laalu (2) :	*unse*]
	Tr :	Retain	

Table 3: Analysis for (6a)

b. jaise $\mathbf{ve}_{1/\#2}$ apnii partii_3 ke taaranhaar kaa svaagat kar rahe
as-if $\text{he}_{1/\#2}$ SELF party_3 of best-man of reception doing was
hon
be-subjunc.
'as if he was receiving the best man of his party.'

(i)	Cb :	Director (1):	*ve*	(ii)	Cb :	Laalu (2) :	*ve*
	Cf :	[Director (1) :	*ve*		Cf :	[Laalu (2) :	*ve*
		party (3) :	*partii*]			party (3) :	*partii*]
	Tr :	Sm-Shift			Tr :	Continue	
		Preferred				**Dispreferred**	

Table 4: Analysis for (6b)

(6a) has a Retain transition, where the pronoun *unse* (which refers to a man called Laalu mentioned in the utterance before (6a)) is the *Cb*. In (6b), the pronoun *ve* can refer to both *Congress director* and *Laalu*, and Step 2 of the BFP algorithm yields a Smooth-shift and a Continue for these two anchors, shown in tables 3 and 4. Step 3 of the BFP algorithm would then rank the *Continue* transition above the *Smooth-Shift*, thus assigning *Laalu* as the antecedent for the pronoun. However, it is the *Smooth-shift* transition which gives the correct and more natural interpretation.[15] Since the use of Rule 2 in the BFP algorithm does not capture the clear preference for the $Cp(U_{i-1})$ in such cases, we propose that the BFP algorithm should be reduced to a simple look-up in the *Cf*-list, the order of which gives the preference for the antecedents of pronouns. This, as will be shown in the next section, is possible with the S-list algorithm since it does not use the centering transitions to compute the antecedents for the pronouns.

[15]The Smooth-shift transition would have been obtained even if the pronoun had been zero instead of overt. This is different from what has been said about Italian by Di Eugenio (1998). In Italian, a Smooth-shift or a Retain is preferred with overt pronouns, but a Continue transition is preferred with null pronouns.

The second type of error generates ambiguities for U_i when the following two conditions hold:

1. (a) when the $Cb(U_{i-1})$ is undefined (after segment boundaries and after intervening utterances without anaphoric relationships to the immediately previous context), or
   (b) when the pronoun under consideratiion cannot co-specify the $Cb(U_{i-1})$, and
2. U_i contains a pronoun (with features not identical to any other pronoun in U_i) which has more than one possible antecedent in $Cf(U_{i-1})$.

Under these conditions the algorithm generates ambiguous readings with the same transition and which cannot be disambiguated by Step 3 of the BFP algorithm. This leads to an ambiguity, and possibly an error chain that could continue throughout the discourse segment and beyond. Condition (1a), where the $Cb(U_{i-1})$ is undefined, is illustrated in example (7).[16] Though the example is from Hindi, such ambiguities would be generated for any language (provided the conditions described in (1) and (2) above hold).

(7) a. B.Singh$_1$ apni$_1$ aadhi se adhik sampatti$_2$ vakilon$_3$ ko bhent
B.Singh$_1$ his$_1$ half than more wealth$_2$ lawyers$_3$ to present
kar chuke the
do perf. had.
'B.SiNgh had presented more than half of his wealth to lawyers.'

(i)	Cb :	none	
	Cf :	[BS (1) :	*B.Singh*
		Wealth (2) :	*sampatti*
		Lawyers (3) :	*vakiilon*]
	Tr :	none	

Table 5: Analysis for (7a)

b. **unki**$_{1/3}$ vartamaan aaya$_4$ ek hazaar rupaye$_5$ vaarshik
his$_1$/their$_3$ present salary$_4$ one thousand rupees$_5$ annually
se adhik na thii.
than more not was.
'His/Their current salary was not more than one thousand rupees annually.'

[16] From "Bare Ghar kii Betii". Short story in *Premchand: Pratinidhi kahaaniyaan.* 1987, p.62.

(i)	Cb :	BS (1):	*unki*	(ii)	Cb :	Lawyers (3) :	*unki*
	Cf :	[BS (1) :	*unki*		Cf :	[Lawyers (3) :	*unki*
		Salary (4) :	*aayaa*			Salary (4) :	*aayaa*
		Rupees (5) :	*rupaye*]			Rupees (5) :	*rupaye*]
	Tr :	Continue			Tr :	Continue	

Table 6: Analysis for (7b)

The analysis for the utterances is provided below the examples. (7a) has no *Cb* and its *Cf*-list has three elements. (7b) contains one pronoun, *unki*. Step 1 of the BFP algorithm generates three anchors for each element of the *Cf*(7a)-list. Step 2 eliminates *wealth* as a possible antecedent because it does not pass the filter (*wealth* is singular whereas the pronoun *unki* is plural/honorific). *BS* and *lawyers* pass the filter since the former is honorific and the latter plural. These two anchors, with *BS* and *lawyers* resolved to the pronoun, are shown in 7b(i) and 7b(ii). Now, Step 3 is applied to rank the transitions for these anchors. However, this cannot be done since both the transitions are Continue (*Cb*(7a) is undefined, and *Cb*(7b) = *Cp*(7b)). Inability to rank the two must leave the pronoun resolved to both *BS* and *lawyers*, thus creating an ambiguity. Cases such as these suggest that the ambiguities are generated because of the use of the *Cb* for the computation of pronominal antecedents. We note again that the antecedent can be correctly selected by a simple look-up in the Cf-list, as is done in the S-list algorithm.

5 The S-list Algorithm

The S-list algorithm (Strube 1998) is based on a model which consists of a single construct, called the S-list, and one operation, the insertion operation. The model is designed to be applied incrementally and describes the attentional state of the hearer at any point in the discourse. The S-list contains some discourse entities which are realized in the current as well as the previous utterance. A ranking is imposed on the elements of the S-list, being determined by information status and/or word order (Strube and Hahn 1999), and the order among the elements provides straightforward preferences for the antecedents of pronominal expressions. However, in Hindi, as we hope to have shown conclusively in Section 3, information status or word order do not seem to affect the salience of discourse entities. Based on our results, we propose the following conventions for ranking the S-list elements in Hindi: the 3-tuple (x, utt_x, gr_x) denotes a discourse entity x which is *evoked* in ut-

terance utt_x with the grammatical role gr_x. With respect to any two discourse entities (x, utt_x, gr_x) and (y, utt_y, gr_y), with utt_x and utt_y specifying the current utterance U_i or the preceding utterance U_{i-1}, we set up the following ordering constraints on elements in the S-list (Table 7).[17]

(1)	If $gr_x > gr_y$, then $x < y$.
(2)	If $gr_x = gr_y$ then if utt_x follows utt_y, then $x < y$, if $utt_x = utt_y$ and pos_x precedes pos_y, then $x < y$.

Table 7: Ranking Constraints on the S-list

The algorithm proposed in Strube (1998) together with the language specific ordering constraints proposed for Hindi resolves the pronouns by a simple look-up in the S-list, and the elements are tested in the given order until one test succeeds. The algorithm (taken from Strube 1998) is given as follows:

1. If a referring expression is encountered,
   (a) if it is a pronoun, test the elements of the S-list in the given order until the test succeeds;[18]
   (b) update S-list; the position of the referring expression under consideration is determined by the S-list-ranking criteria which are used as an insertion algorithm.
2. If the analysis of utterance U is finished, remove all discourse entities from the S-list, which are not *realized* in U.

6 Empirical Data

In this section, we present the results of the application of the BFP algorithm and the S-list algorithm to pronoun resolution in Hindi texts. We used the following guidelines for our tests (see Walker 1989). The basic unit for which the centering data structures are generated is the utterance. The utterance U

[17]The relation $<$ between two entities x and y denotes their relative ordering in the S-list. The relations $>$ and $=$ between gr_x and gr_y indicate that the grammatical role of x is higher then that of y in the ranking heirarchy or that the grammatical roles of x and y are the same.

[18]Testing the elements of the S-list involves checking for agreement features, coreference restrictions and sortal constraints.

is defined as a sentence. Coordinated clauses are taken as separate utterances. A segment is defined as a paragraph unless its first sentence has a pronoun in subject position or a pronoun where none of the preceding sentence-internal noun phrases matches its syntactic features (cf. Walker 1989).

According to the preference for inter-sentential candidates in Centering theory, we defined the following anaphora resolution strategy for the BFP algorithm (at the beginning of a discourse segment the order of steps 1 and 2 is reversed):

1. (a) test elements of $Cf(U_{i-1})$,
2. (b) test elements of U_i left-to-right,
3. (c) test elements of $Cf(U_{i-2})$, $Cf(U_{i-3})$, ...

6.1 Analysis and Results

For our evaluation of the two algorithms, we analyzed some Hindi texts.[19] The results of our analysis are given in Table 8. The first two rows give the number of utterances in the test set and the number of pronouns. The remainder of the table is divided into two parts, each containing results for the two algorithms, respectively. For each algorithm, the numbers of correct and incorrect resolutions are given in the rows marked *correct* and *wrong*. The wrong resolutions are further broken up by the type of error and are described as follows:

- *wrong (strategic)* means that the errors are directly produced by the strategy of the algorithm;
- *wrong (ambiguity)* gives the number of ambiguous analyses;
- *wrong (intra-sentential)* means that the errors are caused by unspecified preferences for intra-sentential antecedents;
- *wrong (chain)* means that the errors were caused by error chains;
- *wrong (other)* gives the remaining errors (for example, errors relating to missing specifications for anaphora across segment boundaries).

[19] The test set consisted of two short stories by Indian novelists, Munshi Premchand and Usha Priyamvada, and one article from a news magazine, *India Today*.

		Prem	Vaap	IT	Total
Utterances		57	139	71	267
Pronouns		41	100	45	186
BFP Alg.	**Correct**	18	79	38	135
	Wrong	23	21	7	51
	Wrong (strategic)	6	5	3	14
	Wrong (ambiguity)	5	2	1	8
	Wrong (intra-sentential)	3	4	1	8
	Wrong (other)	3	5	0	8
	Wrong (chain)	6	5	2	13
S-list Alg.	**Correct**	34	88	44	166
	Wrong	7	12	1	20
	Wrong (strategic)	2	1	1	4
	Wrong (other)	3	5	0	8
	Wrong (chain)	2	6	0	8

Table 8: Results of the BFP and S-list Algorithm

6.2 Evaluation

The *wrong (other)* errors for both algorithms were caused by underspecification for the definition of the discourse segment. In other words, the pronouns were found to select an antecedent too far back in the discourse.[20] The table shows that the BFP algorithm generates errors due to ambiguities (*wrong (ambiguity)*). This is because the algorithm implements the model by comparing possible transitions, which results in inevitable ambiguities in the two types of cases discussed in Section 4. The S-list algorithm, on the other hand, does not generate any ambiguities because of its simple look-up in the S-list for the first possible antecedent match. The BFP algorithm also generates errors due to unspecified preferences for intra-sentential anaphora (*wrong (intra-sentential)*), which are not found in the application of the S-list algorithm because it integrates preferences for inter- and intra-sentential anaphora by making the S-list span across multiple utterances, the current and the previous.

To summarize, our results show (confirming Strube's (1998) results) that the S-list algorithm performs better than the BFP algorithm in general. In particular, we have illustrated that, for a language like Hindi, in which the

[20] We do not pursue this issue here, primarily because of the absence of any precise and implementable definition of the discourse segment (but see Grosz and Sidner (1986)).

ranking is determined by grammatical role and not by information status or word order, the algorithm can be applied straightforwardly if the S-list ranking is made language-specific.

7 Conclusions

In this paper, we proposed a novel method for determining the relative salience in discourse entities and applied this method to Hindi. We concluded that the Cf-list ranking in Hindi is crucially determined by *grammatical role*, and that information status and word order do not have any independent effect on salience. We also applied the proposed ranking to two pronoun resolution algorithms, the BFP algorithm and the S-list algorithm, both of which use the notion of the Cf-list for computing pronominal antecedents, and showed that better results are obtained with an algorithm that does not make straightforward use of the Centering notions of the Cb and the transitions.

References

Brennan, Susan E., Marilyn W. Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. In *Proc. of the 25th Annual Meeting of the Association for Computational Linguistics; Stanford, Cal., 6–9 July 1987*, 155–162.

Chomsky, Noam. 1981. *Lectures on government and binding*. Dordrecht: Foris.

Di Eugenio, Barbara. 1998. Centering in Italian. In *Centering theory in discourse*, ed. M.A. Walker, A.K. Joshi, and E.F. Prince, 115–137. Oxford, UK: Oxford Univ. Pr.

Dimitriadis, Alexis. 1996. When pro-drop languages don't: Overt pronominal subjects and pragmatic inference. In *Chicago Linguistic Society*, volume 32.

Gambhir, Vijay. 1981. Syntactic restrictions and discourse functions of word order in standard hindi. Doctoral Dissertation, Univ. of Pennsylvania, Philadelphia, Penn.

Gordon, Peter C., Barbara J. Grosz, and Laura A. Gilliom. 1993. Pronouns, names, and the centering of attention in discourse. *Cognitive Science* 17:311–347.

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21:203–225.

Grosz, Barbara J., and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics* 12:175–204.

Hoffman, Beryl. 1998. Word order, information structure, and centering in Turkish. In *Centering in Discourse*, ed. M.A. Walker, A.K. Joshi, and E.F. Prince, 251–271. Oxford, UK: Oxford Univ. Pr.

Horn, Laurence. R. 1986. Presupposition, theme, and variations. In *Chicago Linguistic Society*, volume 22, 168–192.

Hudson-D'Zmura, Susan, and Michael K. Tanenhaus. 1998. Assigning antecedents to ambiguous pronouns: The role of the center of attention as the default assignment. In *Centering in discourse*, ed. M.A. Walker, A.K. Joshi, and E.F. Prince, 199–226. Oxford, UK: Oxford Univ. Pr.

Jaeggli, Osvaldo, and Kenneth J. Safir, ed. 1989. *The null subject parameter*. Dordrecht, The Netherlands: Kluwer.

Jaspars, Jan, and Megumi Kameyama. 1995. Discourse preferences in dynamic semantics. In *Proceedings of the Tenth Amsterdam Colloquium*, 445–464.

Kameyama, Megumi. 1985. Zero anaphora: The case of japanese. Doctoral Dissertation, Stanford University, Linguistics Department, Stanford, Cal.

Prince, Ellen F. 1981. Toward a taxonomy of given-new information. In *Radical pragmatics*, ed. P. Cole, 223–255. New York, N.Y.: Academic Press.

Prince, Ellen F. 1992. The ZPG letter: Subjects, definiteness, and information-status. In *Discourse description. diverse linguistic analyses of a fund-raising text*, ed. W.C. Mann and S.A. Thompson, 295–325. Amsterdam: John Benjamins.

Prince, Ellen F. 1994. Subject pro-drop in Yiddish. In *Focus and natural language processing. volume i. intonation and syntax*, ed. Bosch P. and van der Sandt R., 159–173. Working Papers of the Institute for Logic and Linguistics, IBM Deutschland.

Rambow, Owen. 1993. Pragmatic aspects of scrambling and topicalization in German. In *Workshop on Centering Theory in Naturally-Occurring Discourse. Institute for Research in Cognitive Science (IRCS); Philadelphia, Penn.: Univ. of Pennsylvania, May 1993.*

Reinhart, Tanya. 1981. Pragmatics and linguistics. An analysis of sentence topics. *Philosphica* 27:53–94.

Sidner, Candace L. 1979. Towards a computational theory of definite anaphora comprehension in english. Technical Report AI-Memo 537, Massachusetts Institute of Technology, AI Lab, Cambridge, Mass.

Sidner, Candace L. 1981. Focusing for interpretation of pronouns. *American Journal of Computational Linguistics* 7:217–231.

Strube, Michael. 1998. Never look back: An alternative to centering. In *Proc. of the 17th Int. Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics; Montréal, Québec, Canada, 10–14 August 1998*, 1251–1257.

Strube, Michael, and Udo Hahn. 1999. Funtional Centering: Grounding referential coherence in information structure. *Computational Linguistics* 25:309–344.

Turan, Umit. 1995. Null vs. overt subjects in Turkish: A centering approach. Doctoral Dissertation, University of Pennsylvania, Philadelphia, Penn.

Vallduví, Enric. 1990. The informational component. Doctoral Dissertation, University of Pennsylvania.

Walker, Marilyn A. 1989. Evaluating discourse processing algorithms. In *Proc. of the 27th Annual Meeting of the Association for Computational Linguistics; Vancouver, B.C., Canada, 26–29 June 1989*, 251–261.

Walker, Marilyn A., Masayo Iida, and Sharon Cote. 1994. Japanese discourse and the process of centering. *Computational Linguistics* 20:193–233.

Webber, Bonnie Lynn. 1978. A formal approach to discourse anaphora. Doctoral Dissertation, Harvard University.

Rashmi Prasad
Department of Linguistics
University of Pennsylvania
Philadelphia, PA 19104
rjprasad@babel.ling.upenn.edu

Michael Strube
European Media Laboratory GmbH
Villa Bosch
Schloss-Wolfsbrunnenweg 33
69118 Heidelberg
Germany
Michael.Strube@eml.villa-bosch.de